An Interpretable Multi-Modal Transformer-Based Intrusion Detection System Utilizing Log Messages and PCAP Files

Nadia Niknami[1], *Vahid Mahzoon*[2], *Rajorshi Biswas*[3], *Slobadan Vucetic*[2], *and Jie Wu*[2]

[1] *Villanova University*
[2] *Temple University*
[3] *Penn State University(Berks)*

Outline

Background & Motivation

Key Inside

Idea Overview

Evaluation & Conclusion

Problem Background

- Challenges in traditional IDS:
  - Signature-based limitations,
  - Unseen attacks,
  - Poor interpretability

- Logs capture system-level events and context.
- PCAP files provide detailed network-level information.

- By fusing them, we get a more holistic understanding of network behavior

Signature-based vs. Deep Learning vs. Ours

- Signature-based → misses zero-day
- CNN/LSTM-based → weak interpretability, ignores semantics
- Transformer fusion (ours) → interpretable and multi-modal

Prior deep models improved accuracy but still face three key issues:

- **Single modality** – using only packets or only logs.
- **Limited interpretability** – no way to explain detections.
- **Data inefficiency** – models need extensive labeled data

Key Insight: Transformers Unify Modalities

Logs tell us *what happened.* PCAP shows *how it happened.*

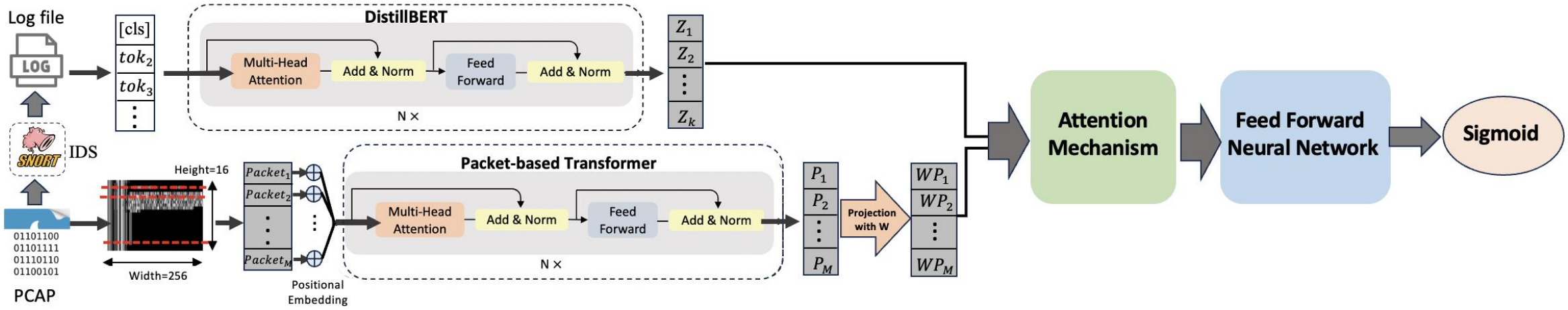
Our idea is to fuse them through a **multi-modal model** so that each modality compensates for the weaknesses of the other

- The **Transformer architecture** was originally designed for text, but later expanded to vision and multimodal tasks.
- We realized this flexibility makes Transformers ideal for merging **packet-level spatial patterns** and **log-level semantic patterns** into one unified framework

Contributions

- A **multi-modal Transformer framework** for IDS
- Integration of **DistilBERT** for semantic log embeddings and a **packet-based Transformer** for traffic data.
- An **attention mechanism** that makes the model interpretable by highlighting important packets or log entries.
- Evaluation demonstrates superior accuracy and generalization to *zero-day attacks*

Idea Overview



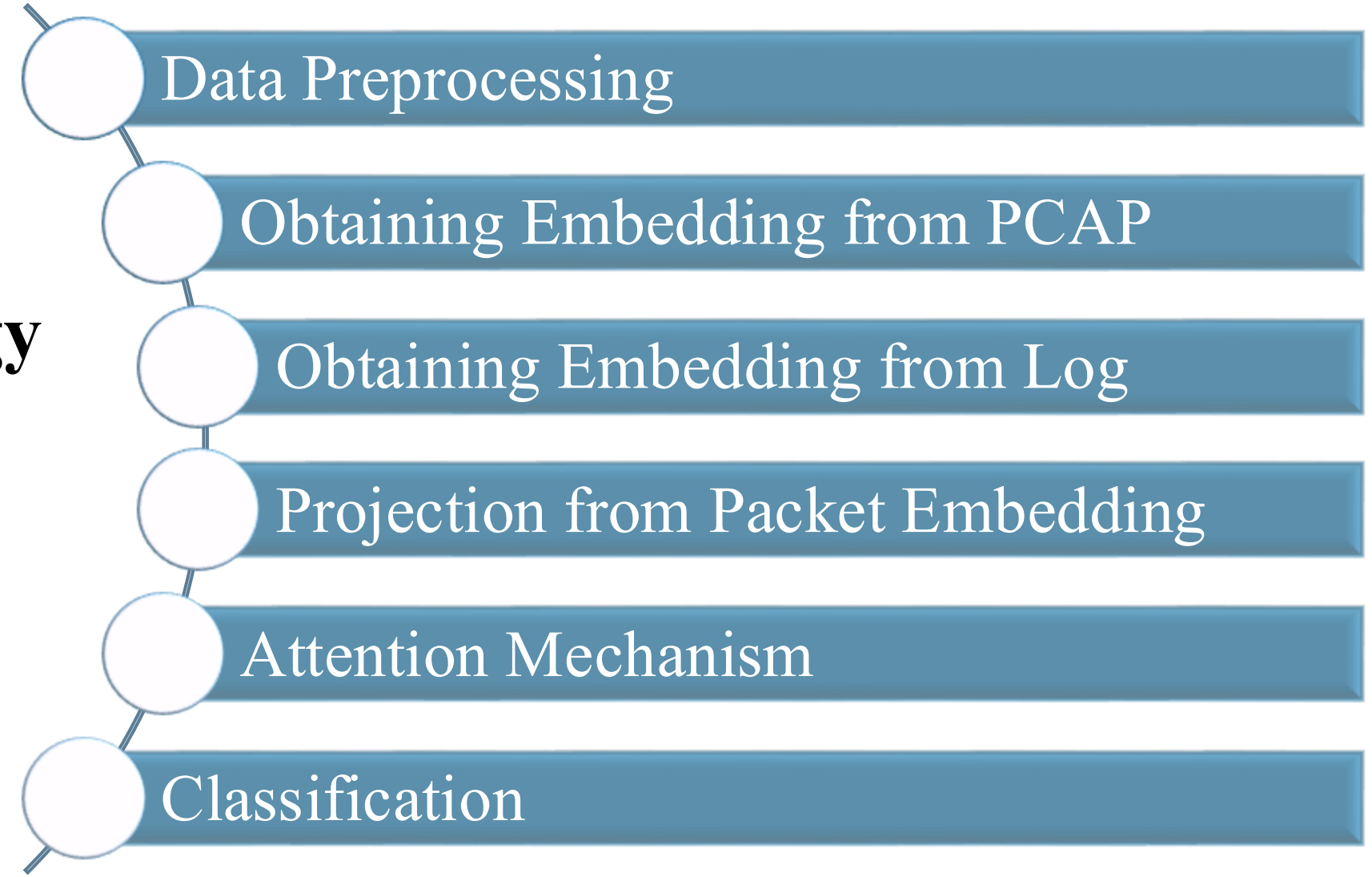
- A **DistilBERT** for logs
- A **packet-based Transformer** for PCAP

The outputs are fused with an **attention module** that learns which modality matters more for each flow

Three goals in bold:

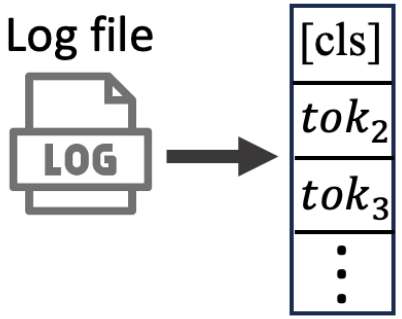
1. Accurate detection (even zero-day)
2. Interpretable results
3. Efficient computation

Methodology of TransIDS

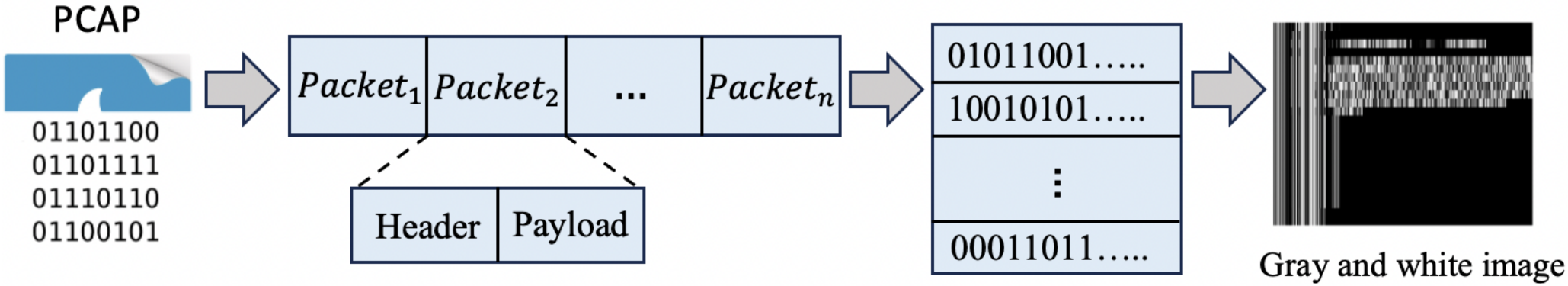


Data Preprocessing

- Log tokenization



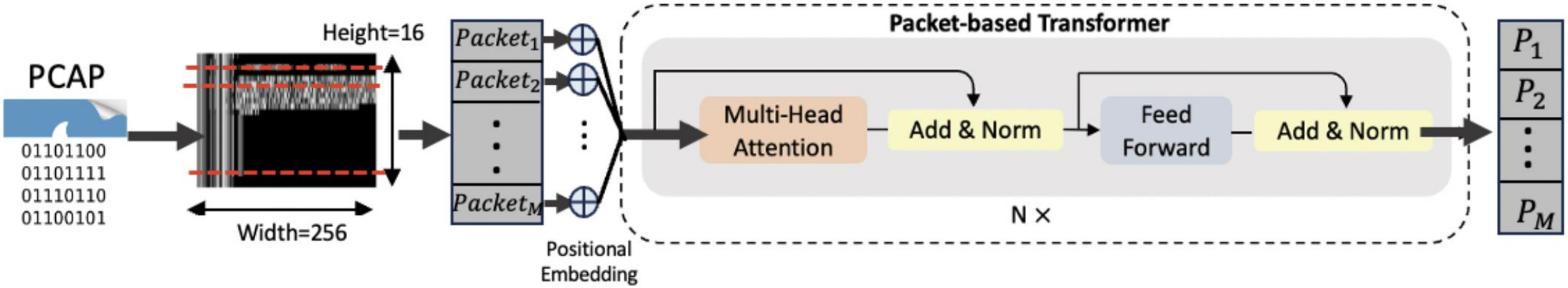
- PCAP-to-image transformation



Obtaining Embedding from PCAP

- The goal of using transformer is to provide the temporal and structural patterns features in the packet data and derive their embeddings.

- This packet-based transformer is inspired by vision transformer.



Obtaining Embedding from Log

- The goal of this module is to provide a vector representation for log content with the help of the BERT model.

- For efficiency, we used a lightweight version of BERT model.

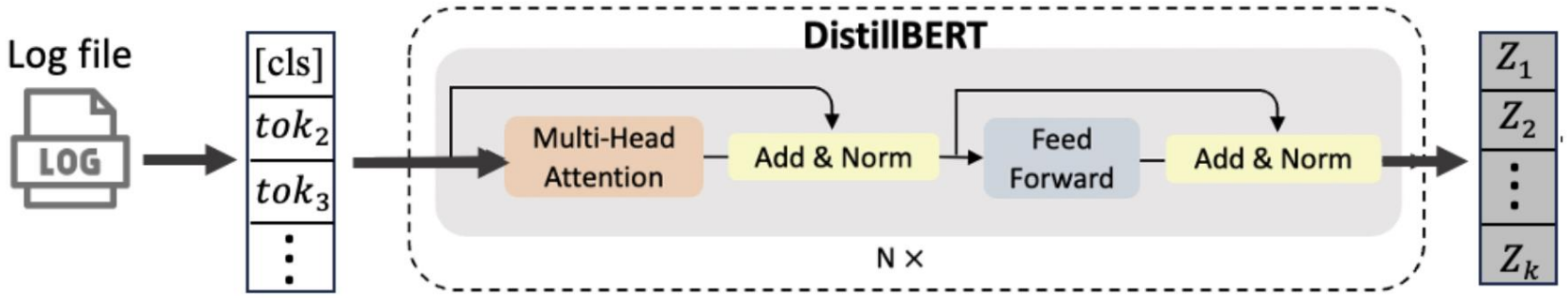
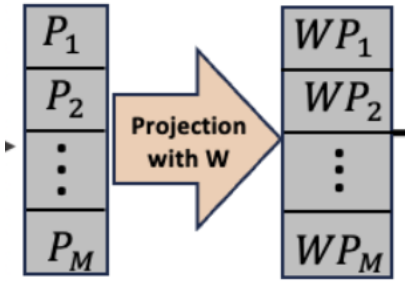


Fig. 3: The Transformer component for analyzing log files in TransIDS framework.

Projection from Packet Embedding

- The obtained embedding for the packets are then projected using a matrix (W) to transform the features into a new space, similar to the embedding size in BERT's embedding space.



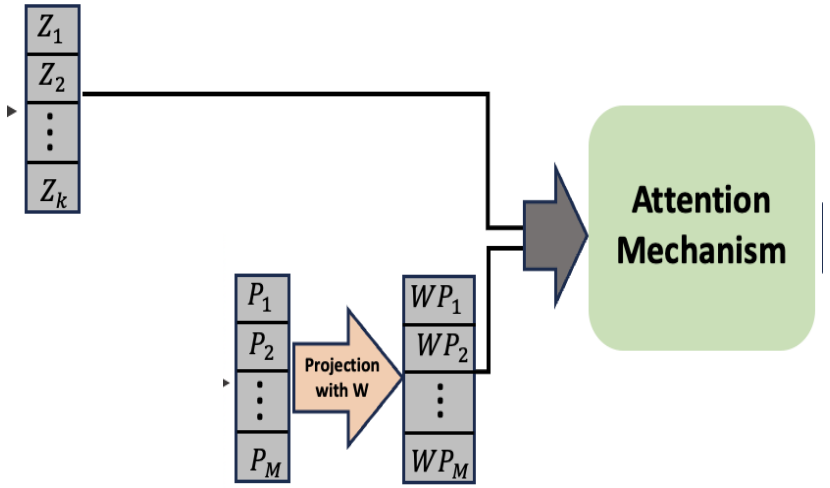
Attention Mechanism

- The attention mechanism is designed to focus on the most relevant parts of the input data when making a prediction.

- It assigns different levels of importance (weights) to various features, packet and log token embeddings, allowing the model to prioritize critical information.

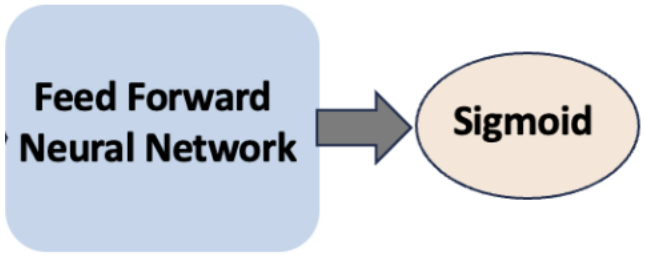
- The output is a single, combined feature set that integrates information from both the log files and the packet data.

- This combined feature set captures the most critical patterns and interactions between the log and packet data.



Classification

- The embedding obtained from the attention mechanism will be passed into a feed-forward neural network with a sigmoid layer to produce the probability that the input data belongs to a certain class.



Evaluation

- CICIDS-2018 dataset, which includes multiple attack categories—DoS, DDoS, BruteForce, and Bot attacks.
- Snort was configured with community and custom rules to generate logs corresponding to the PCAP traffic

TABLE I: Performance metrics of different methods

Method	Individual Classes Accuracy				Overall		
	DDoS	BruteForce	DoS	Bot	Accuracy	Recall	F1-score
DistilBERT	0.7484	0.8119	0.7555	0.6975	0.7534	0.5073	0.6731
CNN-Packet	0.8662	0.9000	0.8950	0.8307	0.8784	0.8458	0.8745
Packet-based Transformer	0.9047	0.9250	0.9175	0.8537	0.9122	0.8776	0.8986
TransIDS	**0.9365**	**0.9700**	**0.9642**	**0.8975**	**0.9424**	**0.8948**	**0.9389**

Evaluation

- Using DistilBERT instead of full BERT cuts inference time nearly in half
- Replacing BERT with DistilBERT maintained identical accuracy but reduced inference time by nearly 50%.

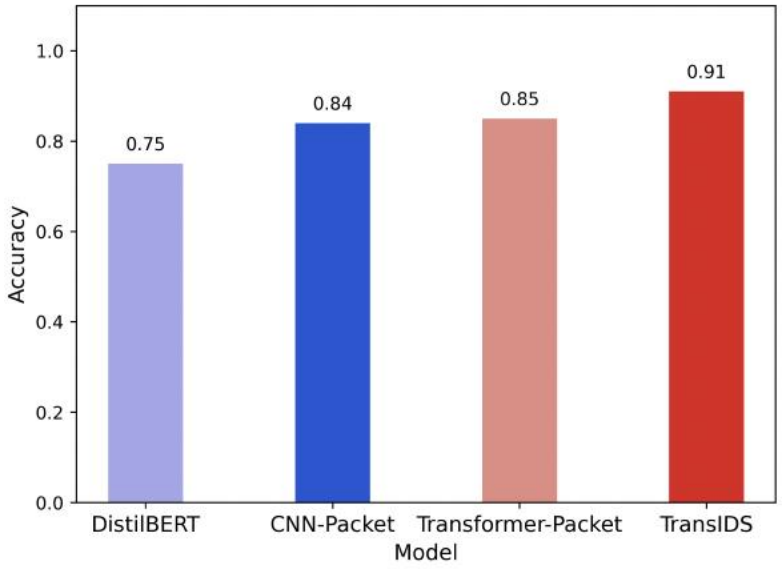


Fig. 4: Multi-class classification.

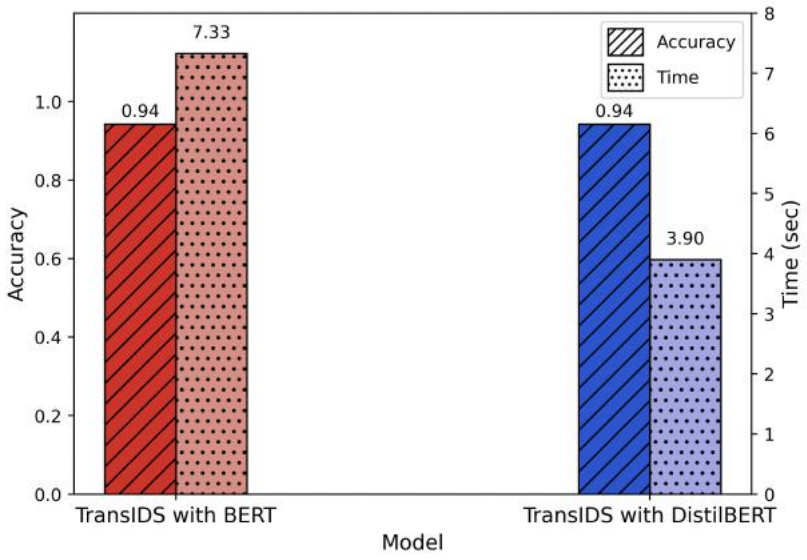


Fig. 5: BERT vs DistilBERT TransIDS.

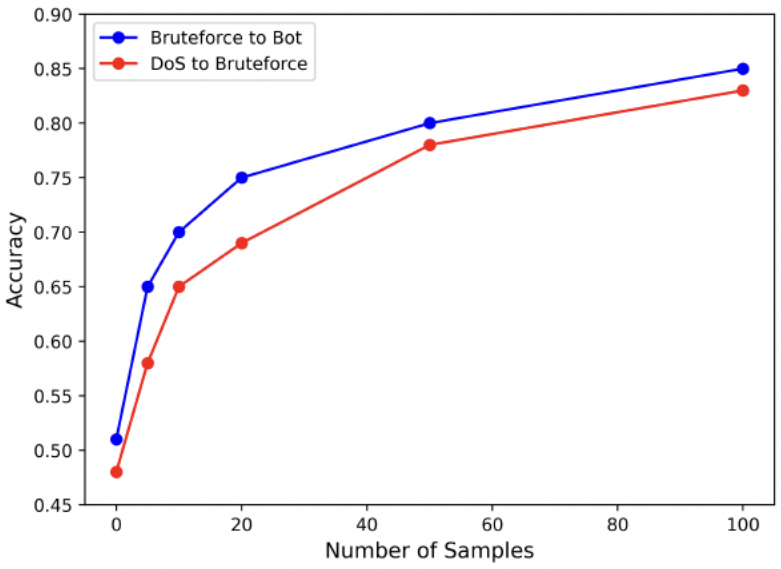
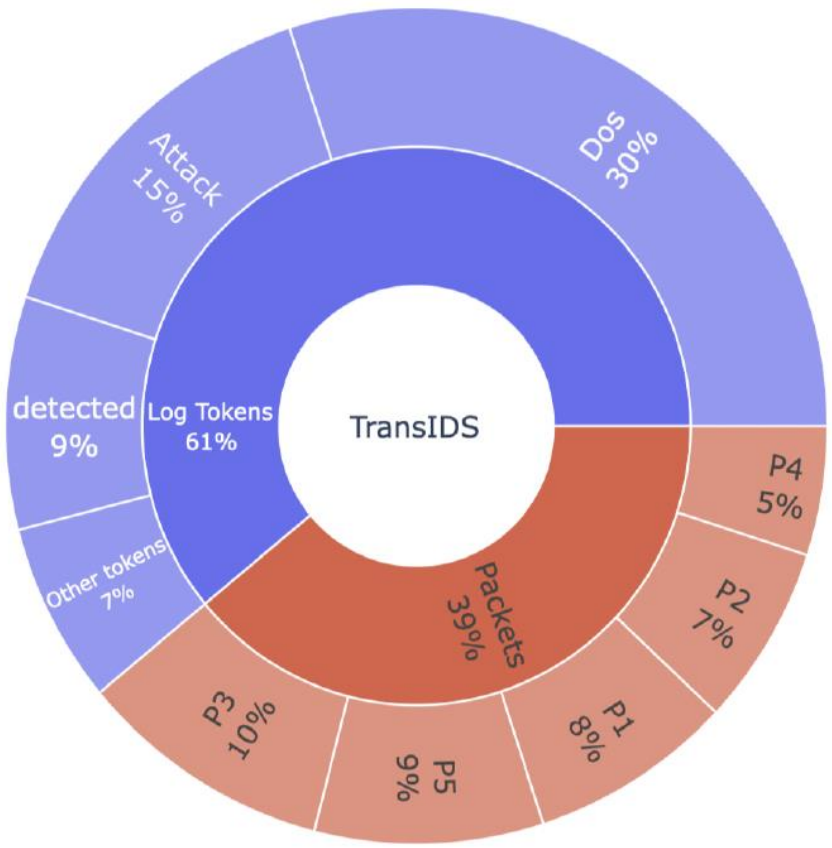


Fig. 6: Few-shot learning.

Evaluation



(a) Flow 1 consists of 5 packets. It is detected by the log-based (signature-based IDS) as "Possible SYN DoS, Possible DoS attack detected".

(b) Flow 2 contains 6 packets. It was not detected by the log-based system and identified as "It is a normal traffic".

Conclusion & Future Work

- **TransIDS** fuses PCAP and log data via Transformers to deliver interpretable, multi-modal intrusion detection.
- It outperforms prior deep models while maintaining transparency.
- Future directions include scaling to larger datasets and applying the framework to *real-time streaming logs* and *cross-domain attacks*

Thank you!

Q & A

nniknami@Villanova.edu